A new formalism for calculation of the partition function of single stranded nucleic acids

Roumen A. Dimitrov
University of Sofia, Faculty of Physics,
Department of Theoretical Physics,
5, James Bouchier Blvd., 1164 Sofia, Bulgaria,
e-mail: dimitrov@phys.uni-sofia.bg

February 9, 2008

Abstract

A new formalism for calculation of the partition function of single stranded nucleic acids is presented. Secondary structures and the topology of structure elements are the level of resolution that is used. The folding model deals with matches, mismatches, symmetric and asymmetric interior loops, stacked pairs in loop and dangling end regions, multi-branched loops, bulges and single base stacking that might exist at duplex ends or at the ends of helices. Calculations on short and long sequences show, that for short oligonucleotides, a duplex formation often displays a two-state transition. However, for longer oligonucleotides, the thermodynamic properties of the single self-folding transition affects the transition nature of the duplex formation, resulting in a population of intermediate hairpin species in the solution. The role of intermediate hairpin species is analyzed in the case when a short oligonucleotides (molecular beacons) have to reliably identify and hybridize to accessible nucleotides within their targeted mRNA sequences. It is shown that the enhanced specificity of the molecular beacons is a result of their constrained conformational flexibility and the all-or-none mechanism of their hybridization to the target sequence.

1 Introduction

Nucleic acids hold great promise as a design medium for the construction of nanoscale devices with novel mechanical or chemical function [1]. Efforts are currently underway in many laboratories to use DNA and RNA molecules for applications in transport, switching [4, 5, 6], circuitry [7], DNA computing [8] and DNA chips [9, 10]. Conformational switches or diversity of conformations have been proven or are suspected to be involved in several important processes such as regulation of gene expression, translational regulation, mutation and repair, and others [11, 2, 14]. During these processes there are several types of interactions trough a network of RNA-RNA, RNA-DNA, RNA(DNA)-protein, RNA(DNA) self-folding or RNA(DNA)- small molecular contacts.

Comparison of short RNAs/DNAs with different base pairs, loop sequences, bulges, etc. has yielded an extremely useful database of thermodynamic parameters from which the stabilities of conformational states of larger nucleic acid sequences can be estimated [3, 19, 20, 21, 22]. The estimation of the thermodynamic parameters is based on nearest-neighbor approximation for inter-residue interactions of closest along the sequence nucleotide residues [23].

There have been several major improvements in calculation of the partition function of a single stranded nucleic acids based on McCaskill algorithm [24, 25, 26] or estimation of the free energy based on free energy minimization and the corresponding sub-ensemble around the minimum free energy conformation [28, 29, 30, 31, 33].

In this work secondary structures and the topology of structure elements are the level of resolution that is used. However, atomic coordinates are also taken into account in the general expressions. Unlike proteins [40], whose secondary structures usually depend on the global amino acid sequence, DNA/RNA molecules are currently thought to assemble in a hierarchical manner [37, 38, 39]. The folding can be conceptually partitioned in the two steps of formation of the secondary structure and the spatial structure [12]. As a result DNA/RNA molecules exhibit a modular structure with individual structural motifs demonstrating independent characteristics.

Therefore, investigation of the overall properties of DNA/RNA molecules based on exploration of variety of local structural motifs, their interactions and distributions along the sequence needs an appropriate theoretical approaches. In particular, this is especially important in a recent increased

interest in predicting target sites for antisense oligonucleotides in highly structured DNA/RNA molecules [41, 44, 13, 42, 43]. Because of the economical value and short experimental cycle, antisense technology has been widly accepted as the tool to study functions of a gene and to validate drug targets. Antisense oligonucleotides can potentially suppress particular gene expression through mechanism such as RNase H-mediated mRNA cleavage, destabilization of the target mRNA or aberation of translation or splicing. Understanding the conformational constraints and transformation between different local structural motifs is of great practical importance. Thus, conformational switches of hairpin-shaped oligonucleotide primers can be useful for enhancing the specificity of nucleic acid amplification reactions. Interactions between short oligonucleotides or small metabolic molecules can lead to conformational switches in the DNA/RNA target molecules [16, 17]. These conformational switches can be used for sensing and modulating complex biochemical networks in variety of important biological processes [15, 14].

Based on such local structural motifs approach in mind, we will use as a starting point our previous work [18], where we presented a new formalism for hybridization processes between DNA and RNA molecules. There hybridization accounted only for stacked pairs, interior loops, bulges and, at the ends, dangling bases. We did not consider stacked pairs in loop and dangling end regions as well as multi-branch loops. The formalism was applied only to short DNA/RNA sequences. Another limitation was that this new formalism was not applied for the estimation of the partition function of self-folding. The self-folding of individual DNA/RNA molecules was based on free energy minimization and the corresponding sub-ensemble around the minimum free energy conformation at each temperature as given by mfold program by Zuker [33]. This led to some inconsistency in the overall calculations. For sequences with non-two state transitions the populations of some intermediate species were poorly predicted. Recently, using McCaskill algorithm [24, 36], mfold has been updated and now it is able to calculate not only the low energy conformations but the ensemble free energy also. It will be interesting in future to compare mfold with the formalism developed here.

In this work we present a new formalism for the estimation of the partition function for self-folding. The formalism use an approach based on the left, right recursion algorithm we have developed for the free energy calculation of duplexes [18]. All possible conformations of single stranded DNA or RNA sequences in solution are explored. The folding model deals with matches,

mismatches, symmetric and asymmetric interior loops, stacked pairs in loop and dangling end regions, multi-branched loops, bulges and single base stacking that might exist at duplex ends or at the ends of helices. Calculations on short and long sequences show, that for short oligonucleotides, a duplex formation often displays a two-state transition. However, for longer oligonucleotides, the thermodynamic properties of the single self-folding transition affects the transition nature of the duplex formation, resulting in a population of intermediate hairpin species in the solution. The advantage of this new formalism is clearly demonstrated especially in the case when one need to design relatively short oligonucleotides (molecular beacons) which have to reliably identify and hybridize to accessible nucleotides within their targeted mRNA sequences. It is shown that the design will enhance the specificity of molecular beacons if they form a stem-and-loop structure with constrained conformational flexibility and an all-or-none mechanism of their hybridization to the target sequence.

2 Methods

2.1 Recursive calculation

With increasing of the temperature the overwhelming majority of the single stranded form conformations tend toward their corresponding unfolded states. At each temperature there is an ensemble of conformational states where each conformation is characterized with the fraction of its base pairs and their location along the sequences which are melted at that given temperature. Thus along the sequences we have variety of local structural motifs characterized by alternating loops -single stranded regions- and double stranded regions. The location and the length of these local structural motifs depend on their relative Boltzmann statistical weights. In this work we are interested to calculate the partition functions of the single-stranded forms based on the method developed for double-stranded forms.

In our previous work (fig.1) [18], the polynucleotide sequences of the double-stranded forms are described as follows: sequence 1 is represented by $S_1 = r_{11}, r_{12}, r_{13}, r_{1i}, r_{1N_1}$ and sequence 2 is represented by $S_2 = r_{21}, r_{22}, r_{23}, r_{2j}, r_{2N_2}$, where N_1 and N_2 stand for their corresponding lengths and r_{1i} and r_{2j} are the space coordinates of the corresponding nucleotides of sequences 1 and 2. The recursion calculation is based on the condition that at least there

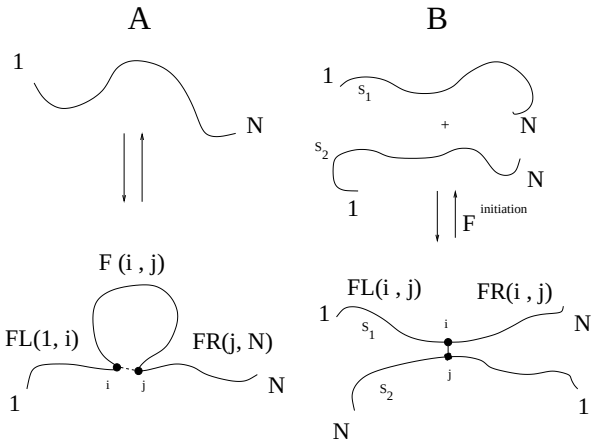


Figure 1: Additive property of the free energy rules based on nearest-neighbor approximation: A- self-folding, B- hybridization [18].

is a two nucleotides along the sequence 1 and sequence 2 that are in contact $r_{1i} - r_{2j}$ and $1 \leq i \leq N_1$, $1 \leq j \leq N_2$. The sequence enumeration is from the $5^{'}$- to the $3^{'}$-end of the sequences. The contact $r_{1i} - r_{2j}$ include an initiation free energy term necessary to bring the two sequences together $F^{initiation}$. Each nucleotide pair $r_{1i} - r_{2j}$ formally divide the hybridized form S_1S_2 of the sequences 1 and 2 in two parts left L and right R in such way that the free energy $F(S_1S_2)$ of S_1S_2 is a sum of the free energies of the left $FL(r_{1i}, r_{2j})$and right $FR(r_{1i}, r_{2j})$ parts plus the initiation free energy $F^{initiation}$ which is assumed to be the same for all possible pairs $r_{1i} - r_{2j}$. Thus,

$$F(S_1S_2) = FL(r_{1i}, r_{2j}) + FR(r_{1i}, r_{2j}) + F^{initiation} \tag{1}$$

This additive property of the energy rules based on nearest neighbor approximation forms the bases of the recursion calculations of the partition function S_1S_2. The additivity of the free energy leads to a multiplication of the partition functions of the left ZL and right ZR parts [18].

Our main focus in this work is the partition function for single-stranded form which similar as we did for the double-stranded form will be described with left and right parts. The sequence is represented by $S = r_1, r_2, r_3, \ldots, r_i, \ldots r_N$, where N stand for it's corresponding length and r_i are the space coordinates of the corresponding nucleotides of sequences S. As previously, the recursion

calculation is based on the condition that at least there is a two nucleotides along the sequence that are in contact $r_i - r_j$.

In contrast to the double-stranded form now the term for the initiation free energy represent the formation of a loop between the positions i and j (fig.1). The sequence enumeration is from the $5^{'}$- to the $3^{'}$-end of the sequence. Each nucleotide pair $r_i - r_j$ formally divide the self-hybridized form of the sequences in three parts left FL, middle FM and right FR in such way that the free energy $F(S)$ of S is a sum of the free energies of the left $FL(r_i)$, middle $FM(r_i, r_j)$ and the right $FR(r_j)$ parts.

$$F(S) = FL(r_1, r_i) + FM(r_i, r_j) + FR(r_j, r_N) \tag{2}$$

The recursion form of the partition functions of the left, middle and right parts have the forms:

Left part:

$$\begin{aligned} ZL(r_1, r_i) &= ZL(r_1, r_{i-1}) + \\ &\quad \sum_{1 \le k < i} ZL(r_1, r_k) \exp\left(-\frac{FM(r_k, r_i)}{RT}\right) \qquad (3) \\ FL(r_1, r_i) &= -RT \ln\left[ZL(r_1, r_i)\right] \qquad (4) \end{aligned}$$

Middle part:

$$\begin{aligned} ZM(r_i, r_j) &= ZM^{open}(r_i, r_j) + \\ &\quad \sum_{i<k<l} \sum_{j>l>k} ZM(r_k, r_l) \exp\left(-\frac{F(r_i, r_j, r_k, r_l)}{RT}\right) \qquad (5) \end{aligned}$$

$$F(r_i, r_j, r_k, r_l) = FL(r_i, r_k) + FR(r_l, r_j) \tag{6}$$

$$FM(r_i, r_j) = -RT \ln\left[ZM(r_i, r_j)\right] \tag{7}$$

Right part:

$$
\begin{aligned}
ZR\left(r_j, r_N\right) &= ZR\left(r_{j+1}, r_N\right) + \\
&\quad \sum_{N \geq k > j} ZR\left(r_k, r_N\right) \exp\left(-\frac{FM\left(r_j, r_k\right)}{RT}\right) \qquad (8) \\
FR\left(r_j, r_N\right) &= -RT \ln\left[ZR(r_j, r_N)\right] \qquad (9)
\end{aligned}
$$

$FL\left(r_1, r_i\right)$ and $FR\left(r_j, r_N\right)$ correspond to the free energy of self-folding of the $5'$ and $3'$ dangle ends of the sequence. Obviously, $FL\left(r_1, r_N\right) = FR\left(r_1, r_N\right)$. The term $FM\left(r_i, r_j\right)$ corresponds to the case of initiation of a loop in the middle part. Thus, $FM^{open}\left(r_i, r_j\right) = -RT \ln[ZM^{open}\left(r_i, r_j\right)]$ represents the free energy initiation of a loop without internal base pair contacts. While, $F\left(r_i, r_j, r_k, r_l\right)$ takes into account the summation over all possible distribution of structural motifs (stack pairs, bulges, symmetric and asymmetric loops, single stranded regions, hairpins and multibranches) along the sequences of the interior regions (i, k) and (l, j). For example when $|k - i| = 1$ and $|l - j| = 1$ the free energy $F\left(r_i, r_j, r_k, r_l\right)$ represents a stack pair which belong to a secondary structure, when $|k - i| = 2$ and $|l - j| = 1$ or $|k - i| = 1$ and $|l - j| = 2$ we have a bulge. When $|k - i| \neq |l - j|$ and there are no any base pair contacts in the loop regions, the free energy $F\left(r_i, r_j, r_k, r_l\right)$ represents an asymmetrical internal loop (including the case of a bulge from the one of the sequences and a loop from the other and another way around), while $|k - i| = |l - j|$ leads to a symmetrical loop (including the case of a bulge from both sequences). The presence of internal base pair contacts in the loop regions lead to hairpins and multibranches. For detailed description of the free energies of the bulges, symmetric and asymmetric internal loops and dangling ends we refer the reader to the recent review by Zuker [35].

And lastly, based on the multiplication property of the partition functions for the left and right parts, for the total partition function we have:

$$
Z\left(S\right) = \sum_{1 \leq i < j \leq N} \left[ZL\left(r_1, r_i\right) ZM\left(r_i, r_j\right) ZR\left(r_j, r_N\right)\right] \qquad (10)
$$

2.1.1 Pair probabilities

Having calculated the partition function will allow us to derive the probability distribution of various conformational properties. However, before that

we need a recursion calculation form for the free energy term $FL\left(r_{1i}, r_{2j}\right)$ in equation (1). This term presents the free energy of the left part in case of hybridization. In our previous work [18] we gave an expression for $FL\left(r_{1i}, r_{2j}\right)$ in which we did not consider stacked pairs in loop and dangling end regions as well as multi-branch loops. Based on our new formalism developed above a general recursion calculation form for the left partition function $ZL^{h}\left(r_i, r_j\right)$ in case of hybridization can be presented as follow:

$$\begin{aligned} ZL^{h}\left(r_i, r_j\right) &= ZL\left(r_1, r_i\right) ZR\left(r_j, r_N\right) + \\ & \sum_{1 \leq k < i} \sum_{N \geq l > j} ZL^{h}\left(r_k, r_l\right) \exp\left(-\frac{F\left(r_i, r_j, r_k, r_l\right)}{RT}\right) \end{aligned} \tag{11}$$

$$FL^{h}\left(r_i, r_j\right) = -RT \ln\left[ZL^{h}\left(r_i, r_j\right)\right] \tag{12}$$

Now we can tern to the calculation of the probabilities of base pairing. For example, the probabilities $P(r_i, r_j)$ and $P(r_i, r_j, r_{i+1}, r_{j-1})$ for single $r_i - r_j$ and double $r_i - r_j, r_{i+1} - r_{j-1}$ base pairs are:

$$P\left(r_i, r_j\right) = \frac{ZL^{h}\left(r_i, r_j\right) ZM\left(r_i, r_j\right)}{Z\left(S\right)} \tag{13}$$

$$P\left(r_i, r_j, r_{i+1}, r_{j-1}\right) = \frac{ZL^{h}\left(r_i, r_j\right) \exp\left(-\frac{F\left(r_i, r_j, r_{i+1}, r_{j-1}\right)}{RT}\right) ZM\left(r_{i+1}, r_{j-1}\right)}{Z\left(S\right)} \tag{14}$$

where $F\left(r_i, r_j, r_{i+1}, r_{j-1}\right)$ is the free energy of base pairing of two nearest-neighbor nucleotides.

Of particular importance is also the ability to monitor the transition between the folded and unfolded structures as well as the partial forms of their conformational intermediates as a function of the temperature by any physical property that is dependent on the number of base pairs formed. Fortunately, the absorption spectra as well as thermodynamics are physical properties that are consistent with the nearest-neighbor models [21, 22]. In other words given nearest neighbors must have identical values of their absorptions or melting free energies regardless of their position in the interior or at the ends of the sequence. In such way the property monitored as a

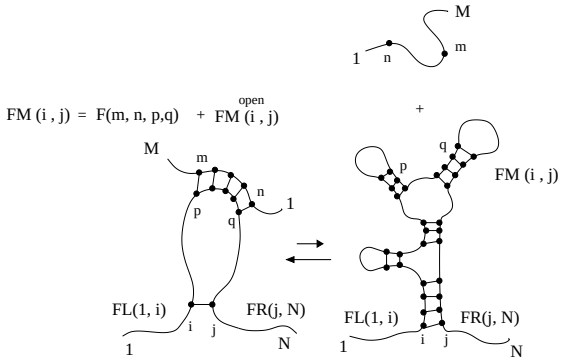


Figure 2: Base pair contacts and their free energy contributions in case of an open loop and branched hairpin. Also an example is given of conformational switching between the loop and the hairpin as a result of interaction of the loop with a short oligo. At the same time the subregion $\{p, , , q\}$ (involved into a multibranched loop) has to unfold before it hybridized with the short oligo.

function of the temperature is proportional to the fraction of base pairs that are stacked as a nucleic acid molecule is melted [18].

Using the base pairing probabilities we can express the equilibrium fraction of bases paired θ as follow:

$$\theta \;\; = \;\; \sum_{ij} P(r_i, r_j) \tag{15}$$

To calculate the extinction we should take into account that it is determined by the contribution of the melted or mismatch loop regions along the constituent sequences of the self-folded species [27]. At each given temperature there is an ensemble of conformation with a narrow or broad distribution of such loops. The contribution of each of them is proportional to its relative Boltzmann statistical weight. It follows from here that the extinction $\epsilon(T)$ for the self-folded species can be represented in the form [18]:

$$\epsilon(T) = \sum_{i=1}^{N-1} 2(1 - P(r_i) - P(r_{i+1}) + P\,(r_i, r_{i+1}))\xi(i, i+1) - \sum_{i=1}^{N-1}(1 - P(r_i))\xi(i) \tag{16}$$

where $1 - P(r_i) - P(r_{i+1}) + P(r_i, r_{i+1})$ is the probability that two closest along the sequence nucleotides with positions i and $i+1$ are melted and as a result give a contribution $\xi(i, i+1)$ to the total absorbance. For the probabilities $P(r_i)$ and $P(r_i, r_{i+1})$ we have:

$$P(r_i) = \sum_{i>n\geq N} P(r_i, r_n) + \sum_{1\leq n<i} P(r_n, r_i)$$

$$\begin{aligned} P(r_i, r_{i+1}) &= \sum_{i+1<n<m} \sum_{n<m\leq N} P(r_i, r_{i+1}, r_m, r_n) + \\ &\sum_{1\leq n<m} \sum_{n<m<i} P(r_i, r_{i+1}, r_m, r_n) + \\ &\sum_{i+1<n\leq N} \sum_{1\leq m<i} P(r_i, r_{i+1}, r_m, r_n) \end{aligned} \quad (17)$$

The formalism developed in this work allow also incorporation of several types of intramolecular interactions trough a network of RNA-RNA, RNA-DNA, RNA(DNA)-protein or RNA(DNA)- small molecular contacts. The additional free energy terms depending on the type of interactions (for example hybridization with short oligos or protein molecules) have to be incorporated into the free energy term $FM(r_i, r_j)$ (fig.2).

3 Results and discussions

Understanding of the molecular forces that control the various sequence- and solvent-specific conformational forms found within DNA and RNA oligonucleotides is of great importance. Melting experiments have been the most useful way to measure variety of thermodynamic parameters from which the stabilities of larger structures under different conditions can be estimated. The estimation of the thermodynamic parameters is based on the assumption that the stability of a base pair is dependent only on the identity of adjacent base pair because the major interactions involved in transformation between different conformations of the polynucleotide sequence are stacking and hydrogen bonding [45, 46, 47, 48]. This additive property of the energy rules based on nearest neighbor approximation forms the bases of the recursion calculations of the partition function. The additivity of the free energy leads to a multiplication of the partition functions [18].

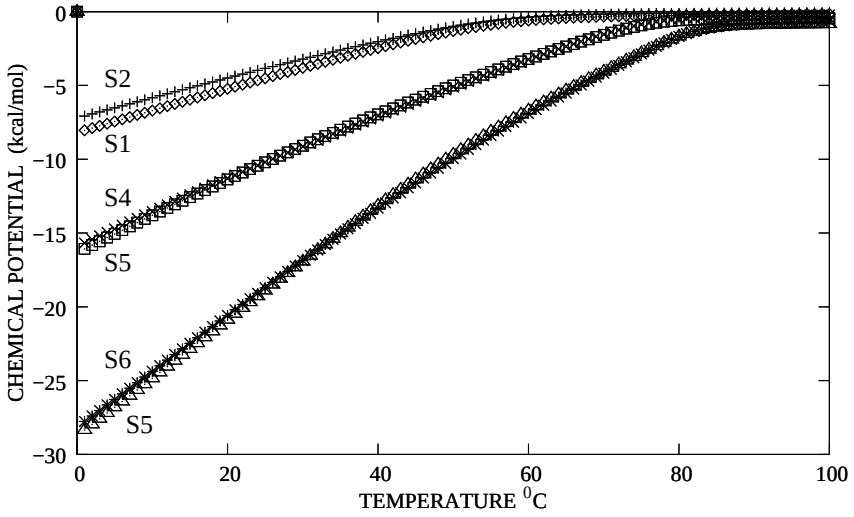


Figure 3: Chemical potential versus temperature for the hairpin species formed after dissociation of the three dsDNAs -S1S2, S3S4, S5S6.

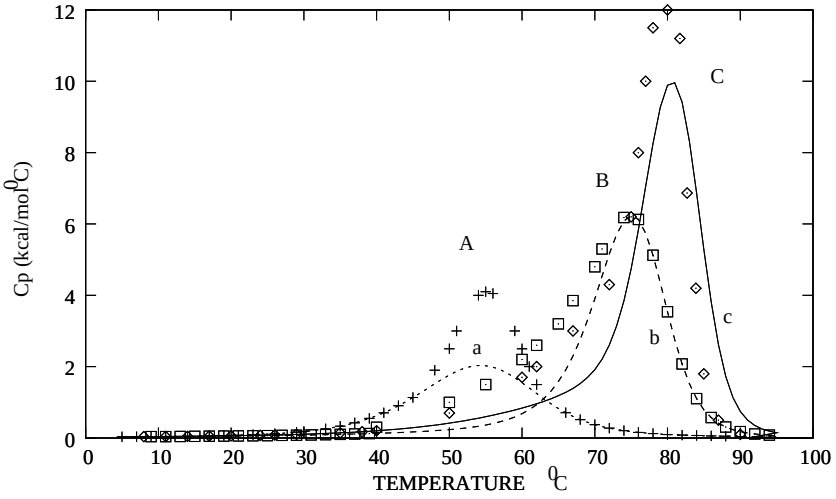


Figure 4: Calorimetric excess heat capacity, ΔC_p, versus temperature profiles for the three dsDNAs. Experimental plots for duplex strand transition are as follows [32]: S1S2(A), S3S4 (B), and S5S6 (C). The calculated curves are with lines and are given as follows: S1S2 (a), S3S4 (b), and S5S6 (c).

Based on the multiplication property of the partition function, here we present a new formalism for calculation of the partition function of a single stranded nucleic acids. The self-folding deal with matches, mismatches, symmetric and asymmetric interior loops, bulges and single base stacking that might exist at duplex ends or at the ends of helices. The formalism also takes into account base pair contacts in the loop regions, or dangle ends in the double helix and single hairpin species as well as multi-branches. This allow calculations of both short and long sequences. The self-folding explores all possible conformations of the single strand species.

We did calculations on non-self-complementary DNA sequences with melting temperatures between 50 C^o and 90 C^o. The sequence length is as follows: 9-S1,d(GCTTGTTGC) and S2,d(GCAACAAGC); 15-S3,d(GCAGGTTGTTTCCGC) and S4,d(GCGGAAACAACCTGC); 21-S5,d(GCAACAGGTTGTTTCCGTTGC) and S6,d(GCAACGGAAACAACCTGTTGC) [32]. The self-folding and hybridization between DNA and RNA sequences takes into account the whole ensemble of single and double strand species in the solution and their fractional extents at different temperatures [18]. We assume that the solution can be described as an ensemble of ideally mixed species. This assumption is based on the experimental evidence that with very good accuracy the single-stranded self-folding trasition and the double-stranded association are independent transition processes and the thermodynamic properties and transition characteristics of each transition in a mixing solution are identical to those in the isolated systems [32]. The calculated chemical potentials of intermadiate hairpin species show that for short oligonucleotides (S1, S2 -fig.3), there is a small thermodynamic contribution of the single-strand self-folding transition to the entire transition. As a result the duplex formation for short oligonucleotides shows a perfectly symmetric two-state shape for the calorimetric excess heat capacity curve versus temperature (fig.4). However, for longer oligonucleotides (S3, S4, S5, S6 -fig.3), calculated chemical potentials show that the thermodynamic properties of the single self-folding transition affect the transition nature of the duplex formation, resulting in a population of intermediate hairpin species in the solution. The deviation of calculated calorimetric excess heat capacity curves versus temperature from a perfectly symmetric shape can be seen for duplexes S3S4 and S5S6 in fig.4. Here, the melting of the intermadiate hairpin species are superimposed on the melting of duplex species thus leading to deviation from the two-state shape of the heat capacity curve.

Further we will analyze in details the transition nature of the duplex

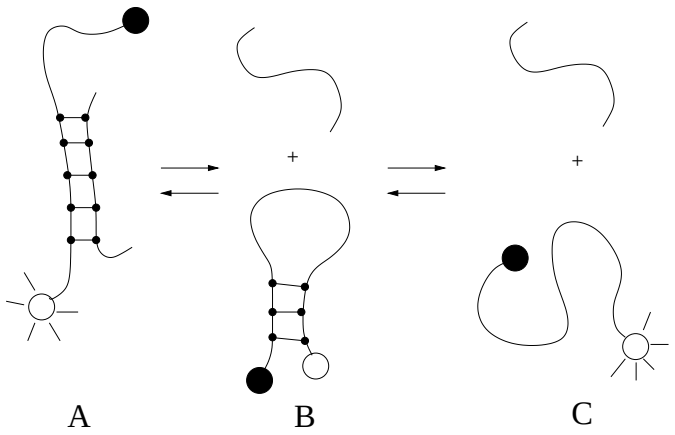


Figure 5: Schematic representation of the phase transitions in solutions containing molecular beacons. At low temperature (phase A) molecular beacons and their targets spontaneously form duplexes. In this state molecular beacons are open and fluorescent. At higher temperature (phase B) duplexes are destabilized and molecular beacons are released, returning to their closed hairpin conformation, and fluorescence decreases. As the temperature is raised further (phase C), the closed molecular beacons melt into fluorescent random coils.

formation or dissociation and the role of the intermediate hairpin species. The role of hairpin intermediates during dissociation or formation of the duplex species in the solution is of great importance in the case when a short oligonucleotides (molecular beacons) have to reliably identify and hybridize to accessible nucleotides within their targeted mRNA sequences. Molecular beacons are DNA probes that form a stem-and-loop intermediate structure and possess an internally quenched fluorophore. When they bind to complementary nucleic acids, they undergo a conformational transition that switches on their fluorescence. Molecular beacons are commonly used to identify complementary strands in the presence of unrelated nucleic acids. Understanding the thermodynamic basis and the underlying conformational transformations of the enhanced specificity of molecular beacons to their target sequences is of great importance. A simple picture based on detailed thermodynamic analysis of the underlying phase transitions in solutions containing molecular beacons is given in fig. 4 [44]. Experimental data give evidence for there phases: phase A- probe-target duplex; phase B- free of target molecular beacon in the form of stem-loop structure and coiled target; and phase C- molecular beacon and the target are both coiled. All-or-none mechanism is supposed for the transitions between the phases. To understand the basis of the molecular beacon specificity from first principle we apply our formalism to calculate variety of thermodynamic characteristics such as free energy, enthalpy and entropy. The idea was to compare the behavior of molecular beacons in the presence of perfectly complementary target oligonucleotides to their behavior in the presence of targets whose sequence created a single mismatched base pair in the probe-target duplex. The sequence of the molecular beacon used in this work is CGCTCCCAAAAAAAAAAACCGAGCG, and the complementary target GGTTTTTTTTTTGG. In our calculations we do not restrict our self to the case of a two-state transitions where in solution during the temperature screening there are only two type of conformational species- fully folded and fully unfolded. Rather we consider the ensemble of all possible intermediate states thus having the most detailed possible picture of the melting process between the folded and unfolded states of the single and double stranded forms. Results from our calculations together with the experimental data are given in Table 1. Our calculations are in very good agreement with the experimental data [44]. Analysis of the calculated melting curves and intermediates, reveals that the enhanced specificity of the molecular beacons is a result of their constrained conformational flexibility and the all-or-none mechanism of their hybridization to the target sequence.

Table 1: Standard enthalpies and standard entropies are shown for solutions containing 50 nM molecular beacons and 1 M target oligonucleotides in the presence of 100 mM KCl and 1 mM $MgCl_2$ [44]. Melting temperatures are for solutions with 50 nM molecular beacons and 300 nM target oligonucleotides. Experiments are given for different mismatches at the same position (marked with 0) and the same mismatch at nearest left (marked with -1) and rigth (marked with +1) positions.

Mismatch	Position	$-\Delta H^0(kcal/mol)$		$-\Delta S^0(eu)$		$T_m(C^0)$	
		exp	cal	exp	cal	exp	cal
T-A	0	84	80	237	238	42	42
A-A	0	69	62	201	202	27	28
C-A	0	61	61.2	175	202	23	28
G-A	0	65	61	185	202	28	28
G-A	-1	72	65	208	218	29	27
G-A	1	74	65	213	217	29	27

Thus, calculations show that the main contribution to the free energy of phase A, in case of perfect match between the probe-target sequences, is practically represented by a single conformational state of the probe-target duplex. The contributions from bulges, interior loops and dangle ends are negligible. The main contributions to the free energy of phase B come from the entropy of the coiled target and the free energy of the loop-stem structure of the molecular beacon. Flexibility of molecular beacon around its hairpin structure is the main way to modulate the stability of phase B. Long stems increase the difference between the melting temperatures of perfectly complementary duplexes and mismatched duplexes. However, too long stems make the hairpin stable not only in phase B but also in phase A. On the other hand, too long hairpin loops decrease the stability of the hairpin. This can lead to disappearance of phase B. Moreover, as the length of the molecular beacon increase, the free energy penalty resulting from a mismatched base pair in the probe-target duplex becomes negligible and will decrease the sensitivity to the presence of a mismatch. Finally, the free energy of phase C is a sum of the entropies of the random coils of both molecular beacon and its target. Our calculations are in full agreement with the experimental data and their thermodynamic analysis (fig. 5)[44].

In conclusion, we presented here a general statistical mechanical approach appropriate to describe the self-folding and hybridization processes of DNA and RNA sequences. The folding model deals with matches, mismatches,

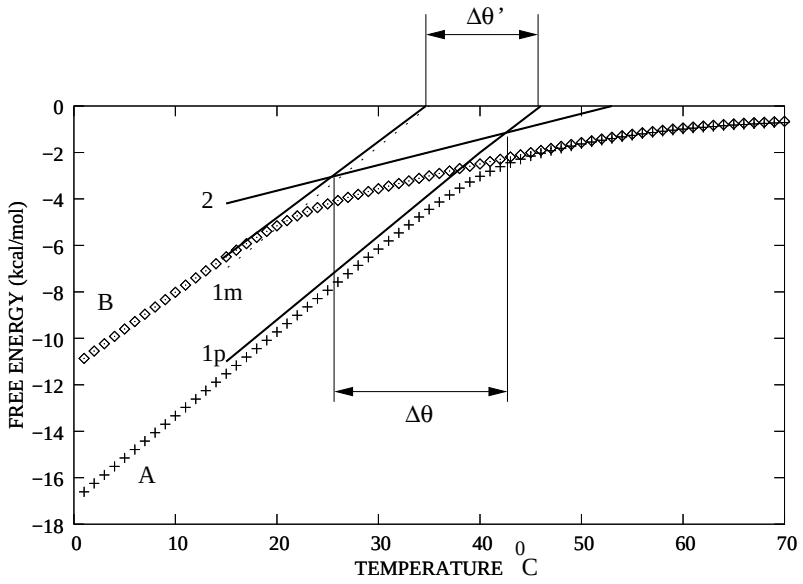


Figure 6: Experimental and calculated free energy of a solution of molecular beacons in equilibrium with target oligonucleotides. Experimental plots [44] for the free energies are as follows: 1p -free energy of the perfect duplex match (phase A); 1m -free energy of the mismatch duplex (phase A); 2 -free energy of the molecular beacon closed form and the coiled target (phase B). The calculated free energy curves are given as follows: A -free energy of the perfect duplex match (phase A); B -free energy of the mismatch duplex (phase A). Since molecular beacons are conformationally more constrained than the unstructured probes, line 2 cross the lines 1p and 1m in such way that increase the difference between the melting temperatures of perfectly complementary duplexes and mismatched duplexes $\Delta\theta$ compare with the $\Delta\theta'$ for an intermediate state of unstructured probe and target.

symmetric and asymmetric interior loops, stacked pairs in loop and dangling end regions, multi-branched loops, bulges and single base stacking that might exist at duplex ends or at the ends of helices. This allow calculations of both short and long sequences.

Calculations on short and long sequences show, that for short oligonucleotides, a duplex formation often displays a two-state transition. However, for longer oligonucleotides, the thermodynamic properties of the single self-folding transition affects the transition nature of the duplex formation, resulting in a population of intermediate hairpin species in the solution. The advantage of this new formalism is clearly demonstrated especially in the case when one need to design relatively short oligonucleotides (molecular beacons) which have to reliably identify and hybridize to accessible nucleotides within their targeted mRNA sequences. It is shown that the design will enhance the specificity of molecular beacons if they form a stem-and-loop structure with constrained conformational flexibility and an all-or-none mechanism of their hybridization to the target sequence. In recent years, a class of diverse regulatory RNAs (often denoted riboregulators) has emerged that regulate expression at the posttranscriptional level. These regulatory RNAs fine tune cellular responses to stress conditions, integrating environmental signals into global regulation. It seems that the structural constraints that enhance the specificity of molecular recognition are also a general feature of the mechanism of action of riboregulators. Thus, the formalism developed in this work can serve as a first step toward creation of a general approach, which can take into account both affinity and specificity of several types of intramolecular interactions trough a network of RNA-RNA, RNA-DNA, RNA(DNA)-protein or RNA(DNA)- small molecular contacts.

References

[1] N. C. Seeman (1999) *Trends Biotechnol.* **17** 437.

[2] S. Gottesman (2002) *GENES and DEVELOPMENT* **16** 2829.

[3] S Freier and D Alkema and A Sinclair and T Neilson and DH Turner (1983) *Biochemistry* **22** 6198.

[4] G. A. Soukup and R. R. Breaker (1999)*Proc. Natl Acad. Sci. USA* **96** 3584.

[5] B. Yurke, A. J. Turberld, A. P. Jr. Mills, F. C. Simmel and J. L. Neumann (2000)*Nature* **406** 605.

[6] H. Yan, X. Zhang, Z. Shen and N. C. Seeman (2002)*Nature* **415** 62.

[7] M. N. Stojanovic and D. Stefanovic (2003) *Nat. Biotechnol.* **21** 1069.

[8] R. S. Braich, N. Chelyapov, C. Johnson, P. W. K. Rothemund and L. Adleman (2002) *Science* **296** 499.

[9] D. D.Shoemaker, D. A. Lashkari, D. Morris, M. Mittman and R. W. Davis (1996) *Nature Genet.* **16** 450.

[10] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan et al. (2000) *Nat. Biotechnol.* **18** 630.

[11] G. Werstuck and M. R. Green (1998) *Science* **282** 296.

[12] I. Jr. Tinoco and C. Bustamante (1999) *J. Mol. Biol.* **293** 271.

[13] D. H. Mathews, M. E. Burkard, S. M. Freier, J. R. Wyatt and D. H. Turner (1999) *RNA* **5** 1458.

[14] G. Stormo (2003) *Molecular Cell* **11** 1419.

[15] M. Mandal, B. Boese, J. E. Barrick, W. C. Winkler, and R. R. Breaker. (2003) *Cell* **113** 577.

[16] M. T. McManus and P. A. Sharp (2002) *Nature Rev. Genet.* **3** 737.

[17] T. A. Vickers, S. Koo, C. F. Bennett, S. T. Crooke, N. M. Dean, and B. F. Baker (2003) *J. Biol. Chem.* **278** 7108.

[18] R. A. Dimitrov and M. Zuker (2003) *Biophysical J.* **87** 215.

[19] N. Sugimoto, R. Kierzek and DH. Turner (1987) *Biochemistry***26** 4554

[20] DR. Hickey and DH. Turner (1985) *Biochemistry* **24** 2086.

[21] JD. Puglisi and IJr. Tinoco (1989) *Methods in Enzymology*, **180** 304.

[22] RD. Blake (1972) *Biopolymers* **11** 913.

[23] PN. Borer, B. Dengler, IJr. Tinoco and OC. Uhlenbeck (1974) *J Mol Biol* **86** 843.

[24] JS McCaskill (1990) in *Biopolymers***29** 1105.

[25] IL Hofacker, W. Fontana, PF. Stadler, S. Bonhoffer, M. Tacker, P. Schuster (1994) *Monatshefte für Chemie* **125** 167.

[26] O. Matzura and A. Wennborg (1996) *Comput Appl Biosci* **12** 247.

[27] C. R. Cantor and I. Jr. Tinoco (1965) *J Mol Biol* **13** 65.

[28] M. Zuker (1989) *Methods Enzymol* **180** 262

[29] AL Williams and IJr Tinoco (1986) *Nucleic Acids Res* **14** 299.

[30] MS Waterman (1983) *Proc Natl Sci USA* **80** 3123.

[31] MS Waterman and TH Byers (1985) *Math Biosci* **77** 179.

[32] P. Wu and N. Sugimoto (2000) *Nucleic Acids Reas* **28** 4762.

[33] M. Zuker (1989) *Science* **244** 48.

[34] M. Zuker (1989) *J. Mol. Biol.* **288** 911.

[35] M. Zuker (2000) *Curr. Opin. Struct. Biol.* **10** 303.

[36] N. R. Markham and M. Zuker (2005) *Nucleic Acids Reas* **33** W577.

[37] R. T. Batey, R. P. Rambo, and J. A. Doudna (1955) *Angew. Chem. Int.* **38** 2326.

[38] E. A. Doherty, R. T. Batey, B. Masquida and J. A. Doudna (2001) *Nature Structural Biology* **8** 339.

[39] T. R. Sosnick and T. Pan (2003) *Current Opinion in Structural Biology* **13** 309.

[40] V. Daggett and A. Fersht (2003) *Nature Rev. Mol. Cell Biol.* **4** 497.

[41] S. P. Walton, G. N. Stephanopoulos, M. L. Yarmush, and C. M. Roth (2002) *Biophysical J.* **82** 366.

[42] S. P. Walton, G. N. Stephanopoulos, M. L. Yarmush, and C. M. Roth (1999) *Biotechnol. Bioeng.* **65** 1.

[43] T. A. Vickers, J. R. Wyaatt and S. M. Freier (2000) *Nucleic Acids Research* **28** 1340.

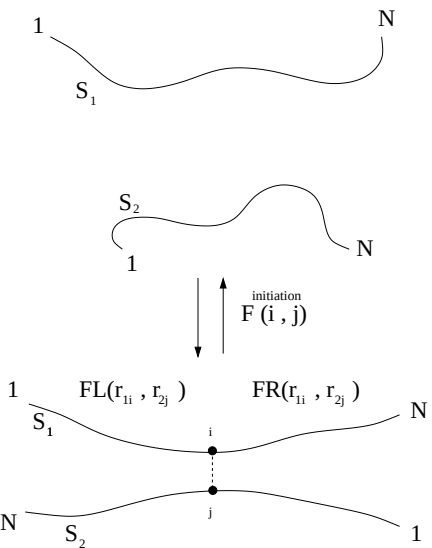
[44] G. Bonnet, S. Tyagi, A. Libchaber, and F. R. Kramer (1998) *Proc. Natl. Acad. Sci. USA* **96** 6171.

[45] S. Freier, D. Alkema, A. Sinclair, T. Neilson, and D. H. Terner (1983) *Biochemistry* **22** 6198.

[46] N. Sugimoto, R. Kierzek and D. H. Terner (1987) *Biochemistry* **26** 4554.

[47] D. R. Hickey and D. H. Terner (1985) *Biochemistry* **24** 2086.

[48] J. D. Puglisi and D. H. Terner (1989) *Methods Enzymology* **180** 304.


1
N
S1
S2
N
1
initiation
F (i , j)
1
FL(r1i , r2j)
FR(r1i , r2j)
N
S1
i
j
N
S2
1